USING DATA MINING TECHNIQUES TO IDENTIFY THE CAUSES OF DEATHS IN AL-GEDAREF HOSPITAL

Ahmed Salah al-Din Abdullah

ABSTRACT: *Data mining technology extensively used in managing relationship through a variety of approaches. There are many tools and methods for analyzing mortality data. The mining technology is one of these tools. The research aims to illustrate the concept of data mining and causes of deaths in Gedaref hospital. The methodology of data mining which used in deaths files is used to integrate two algorithms which are (clustering and classification) to help Gedaref state hospital on prediction and decision making. The study also aims to indicate the level of the interest in the exploration areas and the components of the structure of the application of exploration concepts and tools. One can concluded that the large proportion of deaths is caused by Malaria especially between male's students and employees in early ages 32 year who live in Kassab village in Gedaref. We also recommended that the hospital administration have to provide training programs to workers.*

KEYWORDS: Data Mining, Cause of Death, Database, Computer Application

INTRODUCTION

With the large number of existing data stored in database, this data has become the subject of a question for many researchers to take their advantage. With the increasing prevalence of warehouses or data warehousing, it become necessary to find ways and techniques to extract information and knowledge from such data submitted in solving problems and making decisions using modern computer applications.

Data mining researchers took care of exploration of mortality data by using two algorithms: the classification and clustering. These modern methods of exploration in data mining become the important in solving large complex issues and provide a tremendous amount of alternative solutions during the appropriate time. The solution resulting from the application of the algorithms in the application data is often the ideal solution soon. This method provides the application of intelligent search among a huge number of alternative plans.

Health data is considered the most important forms of knowledge to be excavated, to ensure proper health services and efficiency, and the importance of its role in providing the patient needs in with those hospitals that may be necessary to keep the record for a long time for retrieval according to their types or sources of knowledge obtained.

The analysis of data mining techniques and health operations in Gedaref State Hospital becomes obvious in the belief that the importance of data mining is similar to the importance of data necessary for the implementation of health processes and information. Because of the presence of a huge amount of mortality files without taking advantage of them, the researcher exploring the data to discover the correlation between marital status and age in the disease mortality data to the detect the growing proportion of diseases that have spread and become a formidable deaths and to find the relationship by analyses.

Data mining is often set in the broader context of knowledge discovery in databases, or KDD. This term originated in the artificial intelligence (AI) research field. The KDD process involves several stages: selecting the target data, preprocessing the data, transforming them if necessary, performing data mining to extract patterns and relationships, and then interpreting and assessing the discovered structures.[1]

Data mining process is known as analytical process of exploration and research in the huge and enormous data to extract useful patterns and find relationships and the extent of the correlation between the elements. Data mining usually deal with data have been obtained for another purpose other than the purpose of data mining, for example, database transactions in a bank, which means that the method of data mining is not at all affect the way the of data collection. This is referred to as the process of data mining as secondary statistical process. The definition also indicates that the amount of data is usually large. In case of small amount of data it is preferable to use regular statistical methods for analysis[2, 3, 4,5,6,7,8,9,10].

Saving and organizing information helps to run the business, and to take advantage of mortality information in the evaluation and development of the hospital system, beside it is used as a means of documentation are referenced when needed, has won Data Mining deal of attention in the areas of making information. This is due to the abundance of wide massive amounts of data and the possibility converted into useful information and knowledge, which can be found to extract the data.

The main objective of this research is to build a pattern helps to predict mortality reasons in Al-Gedaref State Hospital using WEKA application and classification and clustering algorithm on the basis of transactions that can be used as a tool in mining to see the proportion of deaths in Al-Gedaref Hospital and learn more about the diseases that cause death for the most influential sectors of society, age, and areas. And also to analyze the data for new unexpected relationships, such as the relationship of the disease in the region. Finally, to see the future forecasts and knowledge discovery, allowing the generation of right decisions to be taken in a timely manner. Storing of information and organization by these techniques help to run the business as it helps to take advantage of mortality information in the evaluation and development of the hospital system, and use it as a means of referenced documentation when needed.

Research interested in exploration in the mortality data of Al-Gedaref Hospital State, one of the states of Sudan, which lies in the eastern part of it, as in the free encyclopedia and Wikipedia[11] between latitudes 12-17 degrees north and longitude 34 36 degrees east longitude. Bounded on the north and west side by the Khartoum and Gezirah state, and on the east is surrounded by Kassala State and Sudanese-Ethiopian border, and Sennar state from south. The number of hospitals where 22, including Al-Gedaref Teaching Hospital, which was founded in 1918 and includes Eye, accidents, ear, nose and throat, children, women and childbirth divisions. There is a department of statistics and information, which was founded in 1969 and is one of the largest sections because it contains all the information pertaining to patients and deaths.

The expected result of this research is to highlight the rules of engagement to see more diseases that cause deaths and at any class of any society ages, more injury and contact data that helps the user to access the decision-making base.

METHODOLOGY

WEKA[12] program gives a range of machine learning algorithms for data mining tasks. The algorithms are applied directly to the data or by using code such as JavaScript. It contains tools for data processing such as classification, regression, assembly, association rules, and visualization. It is well suited for the development of a new learning plans, which is also open source software released under the GNU General Public License.

One way to use WEKA is to apply learning by the way of the data set and analysis of outputs to learn more about the data. It can also use models that have already been taught to generate expectations on the new cases, and also can be used for a variety of models that have already been taught, and compare their performance in order to use it for prediction.

In WEKA interactive interface, one can select the learning mode to be used from the list, and a lot of ways to have a range of disciplinary parameters that can be accessed through the menu or through the object editor. The unit is used to joint assessment to measure the performance of all works. The applications of actual learning systems are the most valuable resources offered by WEKA, but the tools such as filters which are necessary for processing the data comes next. As in the case of classification, one can specify a list of filters and modify them for his requirements.

Often data is displayed in the form of Excel Sheet or database. The original way of storage in WEKA is in the form of Attribute Relation File Format and can easily transfer from software scheduling form to ARFF. Most of the ARFF files consists of a list of cases and the values of the properties of each of the cases that have been separated by a comma. Most of scheduling programs and databases allow the export of data in a file format of Comma Separated Value (CSV) in the form of a set of records with an interval between values. Dealing with the mortality files, one find that a lot of available information is untapped. In this research, a sample of data of 280 records was identified as in Table 1. It has been processed according to the steps of data exploration.

Table 1 a sample of mortality data set of Al-Gadaref hospital

Marital status	City	State	Disease	Job	Age	Sex
Married	Firdos	Al-gadaref	Myocarditis	Business	60	Male
No	Danagla	Al-gadaref	Klazar	Student	13	Male
Married	Kassab	Al-gadaref	Malaria	Labor	45	Male
No	Jamhoria	Al-gadaref	Myocarditis	Student	22	Male
Married -F	Gareeb	Al-gadaref	Kidney failure	Household	38	Female
Married -F	Algoraisha	Al-gadaref	Stroke	Household	75	Female
No	Altadamon	Al-gadaref	dysentry	No	5	Male
No	Basanda	Al-gadaref	Anemia	Student	15	Male
No	Aljinaina	Al-gadaref	Malnutrition	Student	9	Male
Married -F	Alfashaga	Al-gadaref	Blood pressure	Household	50	Female
No	Algoraisha	Al-gadaref	Malaria	Student-F	11	Female
No	Almalik	Al-gadaref	Malaria	Officer	25	Female

It was subsequently modified as in table 2 by changing Married –F to be married, and Student-F

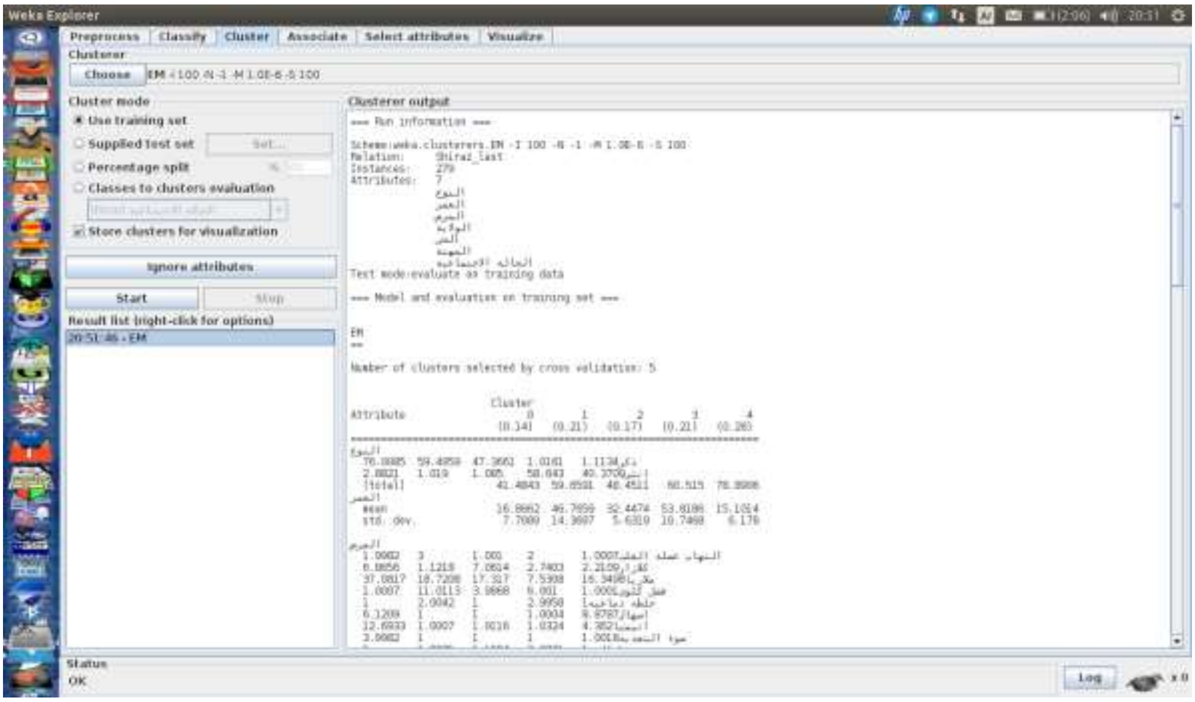
To be Student. Training set has been saved as one worksheet within the Excel program with csv extension, and WEKA program was opened in explorer position. Preprocess tab was clicked, and the previous file with extension csv was chosen and select all the properties of open button. When Classify was chosen and classifier-trees-j48 tab (C4.5) was selected, and test group was selected, and clicking the Start button, the result appears in detail.

Another process starts from data mining operations, a process classification (classify) and its consequences on the sex and found that: ZeroR predicts class value = Male, and found that the value of the most widespread of the disease is malaria and value of the most predictive city was Kassab, and it turned out that the value of predictive job are of the student. On the social side, it turns out that the level of predictive value of the disease is in those without marriage.

Table 2 preprocessing of mortality data in table 1

Marital status	City	State	Disease	Job	Age	Sex
Married	Firdos	Al-gadaref	Myocarditis	Business	60	Male
No	Danagla	Al-gadaref	Klazar	Student	13	Male
Married	Kassab	Al-gadaref	Malaria	Labor	45	Male
No	Jamhoria	Al-gadaref	Myocarditis	Student	22	Male
Married	Gareeb	Al-gadaref	Kidney failure	Household	38	Female
Married	Algoraisha	Al-gadaref	Stroke	Household	75	Female
No	Altadamon	Al-gadaref	Dysentery	No	5	Male
No	Basanda	Al-gadaref	Anemia	Student	15	Male
No	Aljinaina	Al-gadaref	Malnutrition	Student	9	Male
Married	Alfashaga	Al-gadaref	Blood pressure	Household	50	Female
No	Algoraisha	Al-gadaref	Malaria	Student	11	Female
No	Almalik	Al-gadaref	Malaria	Officer	25	Female

From the perception of the screen, it is divided into two types: males and females, and it turns out that unmarried are more infected with malaria and less injury in some diseases as in fig 1.



Clusterer: EM -I 100 -N -1 -M 1.0E-6 -S 100

Cluster mode:
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
☑ Store clusters for visualization

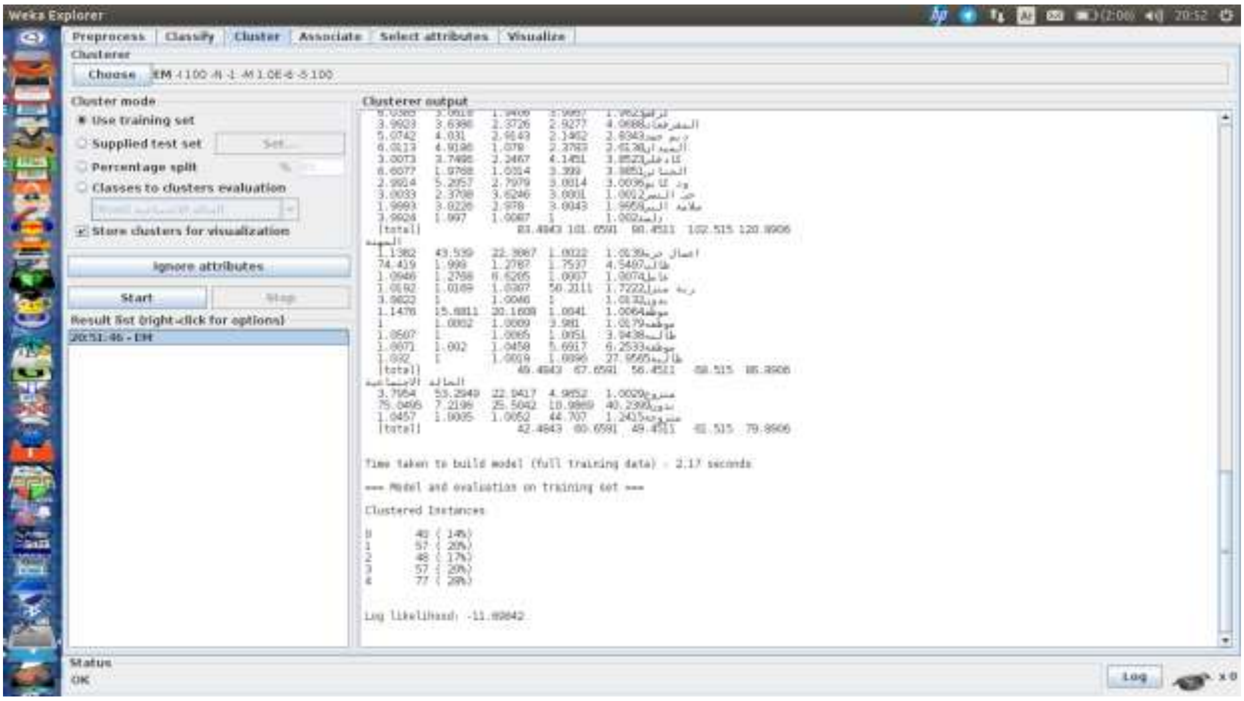
Result list: 20:51:46 - EM

Status: OK

```
=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     Shiraz_last
Instances:    279
Attributes:   7
              النوع
              العمر
              المرض
              الولاية
              الحي
              المهنة
              الحالة الاجتماعية
Test mode:evaluate on training data

=== Model and evaluation on training set ===


EM
==

Number of clusters selected by cross validation: 5


                       Cluster
Attribute                    0        1        2        3        4
                        (0.14)   (0.21)   (0.17)   (0.21)   (0.28)
====================================================================
النوع
  76.8885  59.4859  47.3662  1.0161   1.1134ذكر
  2.8821   1.019    1.005    58.643   40.3709انثى
  [total]                  41.4843  59.6591  48.4511  60.515  78.8906
العمر
  mean                     16.8862  46.7859  32.4474  53.8188  15.1014
  std. dev.                 7.7089  14.3687   5.6319  10.7468   6.176

المرض
  1.0002   3        1.001    2        1.0007التهاب عضلة القلب
  8.8856   1.1218   7.0614   2.7403   2.2109كلازار
  37.0817  18.7208  17.317   7.5308   16.3498ملاريا
  1.0007   11.0113  3.9868   6.001    1.0001فشل كلوي
  1        2.0042   1        2.9958   1جلطة دماغية
  6.1209   1        1        1.0004   8.8787اسهال
  12.6933  1.0007   1.0016   1.0324   4.3821انيميا
  3.0002   1        1        1        1.0018سوء التغذية
```

Fig 1 dividing the data set to cluster

The screen was split according to the ages of in range (5- 47.5 to 90), and those without marriage appeared to be more affected by malaria, and less affect in some other diseases as in fig 2 and fig 3.



Clusterer: EM -I 100 -N -1 -M 1.0E-6 -S 100

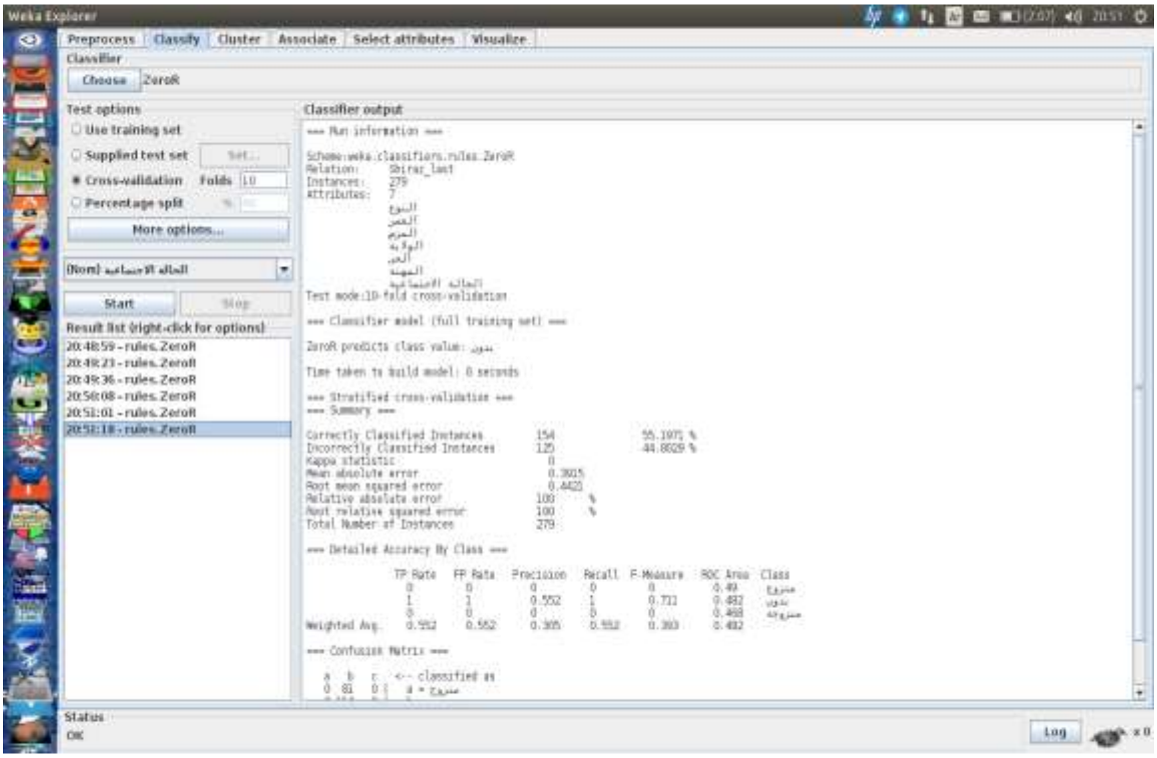
Cluster mode:
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
☑ Store clusters for visualization

Result list: 20:51:46 - EM

Status: OK

```
  3.9923   3.6398   2.3726   2.9277   4.0688[illegible]
  5.0742   4.031    2.9143   2.1962   2.8343[illegible]
  6.0213   4.9198   1.078    2.3783   2.6138[illegible]
  3.0073   3.7498   2.2467   4.1451   3.8521[illegible]
  6.6077   1.9768   1.0314   3.399    3.3851[illegible]
  2.9914   5.2057   2.7979   3.0014   3.0036[illegible]
  3.0033   2.3708   3.6246   3.0001   1.0012[illegible]
  1.9993   3.0228   2.978    3.0043   1.9958[illegible]
  3.9928   1.997    1.0087   1        1.002[illegible]
  [total]                  83.4843 101.6591  90.4511 102.515 120.8906
المهنة
  1.1382   43.539   22.3867  1.0022   1.0139اعمال حرة
  74.419   1.999    1.2787   1.7537   4.5497طالب
  1.0946   1.2758   6.6265   1.0007   1.0074عاطل
  1.0192   1.0169   1.0307   50.2111  1.7222ربة منزل
  3.9822   1        1.0046   1        1.0132بدون
  1.1476   15.8811  20.1608  1.0041   1.0064موظف
  1        1.0002   1.0009   3.981    1.0179موظفه
  1.0507   1        1.0065   1.0051   3.9438طالب
  1.0071   1.002    1.0458   5.6917   6.2533موظفه
  1.002    1        1.0019   1.0096   27.9565طالبه
  [total]                  49.4843  67.6591  56.4511  68.515  85.8906
الحالة الاجتماعية
  3.7954   53.2949  22.9417  4.9852   1.0029متزوج
  75.0495  7.2198   25.5042  10.9869  40.2398بدون
  1.0457   1.0005   1.0052   44.707   1.2415متزوجه
  [total]                  42.4843  60.6591  49.4511  61.515  79.8906


Time taken to build model (full training data) : 2.17 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       40 ( 14%)
1       57 ( 20%)
2       48 ( 17%)
3       57 ( 20%)
4       77 ( 28%)


Log likelihood: -11.89842
```

Fig 2 dividing the data set to cluster (continued)



Classifier: ZeroR

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split

(Nom) الحالة الاجتماعية

Result list (right-click for options)
20:48:59 - rules.ZeroR
20:49:23 - rules.ZeroR
20:49:36 - rules.ZeroR
20:50:08 - rules.ZeroR
20:51:01 - rules.ZeroR
20:51:18 - rules.ZeroR

Classifier output

```
=== Run information ===

Scheme:weka.classifiers.rules.ZeroR
Relation:     Shirar_last
Instances:    279
Attributes:   7
              النوع
              العمر
              المرض
              الولاية
              الحي
              المهنة
              الحالة الاجتماعية
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: بدون

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         154               55.1971 %
Incorrectly Classified Instances       125               44.8029 %
Kappa statistic                          0
Mean absolute error                      0.3915
Root mean squared error                  0.4425
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              279

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0         0          0        0          0.49     متزوج
                 1         1         0.552      1        0.711      0.482    بدون
                 0         0         0          0        0          0.468    متزوجة
Weighted Avg.    0.552     0.552     0.305      0.552    0.393      0.482

=== Confusion Matrix ===

  a   b   c   <-- classified as
  0  81   0 |   a = متزوج
```

Status
OK

Fig 3 classifying the mortality data in marital status

RESULTS AND DISCUSSION

After conducting many experiments on a WEKA program by taking age, job, marital status, city and disease attributes, it was found that there is a relationship between the age and job, marital status, city and disease with each other by applying classification and clustering algorithm.

Clustering and classification algorithm are used together in order to overcome the problem of the large number of deaths. The results showed that when the algorithms merged together, it led to the accurate model added improvement to by classifying deaths.

When the data set was subdivided into male and female, it was found that male matrix values is 0.495 and females matrix is reversible. By dividing the city to matrices, it appeared that malaria is in the district of (Kasab). When rating the job to student's matrices, it was found that they are the most vulnerable to the disease than the rest of other professions. The classification of marital status to matrices has emerged that the values without marriage only susceptible to disease .

Table 3 showed that the groups differ from each other and members are similar to some extent.

Table 3 clustering of the data set in different attributes

Cluster Attribute	1	2	3	4	5
Sex	41.4843	59.6591	48.4511	60.515	97.8906
Age	16.8662	46.7650	32.4474	53.8186	15.1014
Disease	60.4843	78.6591	67.4511	79.515	97.8906
State	40.4843	58.6591	47.4511	77.8900	59.515
City	83.4843	101.6591	90.4511	102.515	120.8906
Job	42.4843	67.6591	56.4511	68.515	86.8906
Marital status	42.4843	60.6591	49.4511	61.515	79.8906
Average	40	57	48	57	77
Percentage	%14	%20	%17	%20	%28

A correlation was also found with the disease and sex when the screen is divided into two types; male and female. For unmarried males understand more infected with malaria, and less injury in certain diseases and lacking in both breast and thyroid tumor and female unmarried women also are more susceptible to malaria and fall injury in certain diseases and lacking in cirrhosis. There is a correlation between the disease and age were divided ages of age (5-47.5 to 90) and appeared to be without a marriage of both sexes more infected with malaria from the age (5 to the prior 47.5) and less in some diseases and have zero stroke. The correlation with disease biology has shown that unmarried men and women are more infected with malaria, all from the countryside. In correlation with disease profession unmarried, the most infected with malaria are a class of students.

In correlation with the social situation of the disease, the most infected with malaria from the category of non-married couples. It is also noted that unmarried males In early age until the age of 32 who are students, staff and entrepreneurship category most infected with malaria, especially in rural Kassab Gedaref area.

RECOMMENDATIONS AND CONCLUSION

The proportion of the increasing number of deaths in Al-Gedaref Hospital, was used for the process of data mining. The lack of sufficient time, and failure to collect sufficient data, a sample of data set of 280 records is used to help in the diagnosis of certain diseases. The researcher found that most of the deaths due to malaria, a disease that affects males at an early age before reaching the age of nineteen. Having touched on one of the techniques used so far, one can recommend the following:

1. The researchers have to use the analysis tools such as: classification and clustering in order to expand the research to include the diagnosis of diseases, and predict the viability of causing the injury of death.
2. The use of new applications will help in the decision-making process.
3. Hospitals collect more data about mortality to be used for more exploration that can be used to increase the efficiency of the results.

REFERENCE

David Hand, Heikki Mannila and Padhraic Smyth, “Principles of Data Mining”, the MIT Press, 2001

Colin Ferguson and Paul Nevell, “The relationship between Machine enjoyment, computer attitude and computer usage: some further refinements”, vol.36 Issue1, pages 113-125, May 1996 accounting.

V. Kumar and K. Chadha, “Mining association rules”, vol.g.no.5, pp.211 – 216, 2012

Graham Williams, “Data mining desktop survival guide”, springer, 2011

Jasmine, et al, “Classification of credit applications using data mining”, 2002.

C. Romero, et al, “Mining raeassociation rules from e-learning data”, pp.171-180

Jaideep Srivastava et al, “Web mining, Accomplishments and Future direction”, Computer Science Department, 200 Union Street SE, 4-192, EE/CSC Building University of Minnesota, Minneapolis, MN 55455, USA.

Rowe and Neil c., “Artificial intelligence through prolog”, prentice hall international Inc. USA, 1988.

S. Cost et al, “Nearest neighbor algorithm for learning with symbolic features machine learning classifier”, ijicis, vol. 1993.

Rekha Bhowmik, “Data Mining Techniques in Fraud Detection”, Journal of Digital Forensics, Security and Law, Vol. 3(2) www. wikipedia.com accessed at 10/10/2015.

Ian H. and Eibe Frank, “WEKA, Machine Learning Algorithms in Java”, Morgan Kaufmann Publishers, 2000